ISSN 0100-6991

WEIGHT OF EVIDENCE

A IMPORTÂNCIA DA EVIDÊNCIA

Alexandre Santos Aguiar[1]; Carlos Alberto de Bragança Pereira[2]

INTRODUCTION

Recording medical information seems to have started just after development of writing around 3100BC[1,2,3,4]. These ancient records were purely descriptive. Although the oldest report of an experiment in humans has taken place around 600BC according to the Old Testament and some key concepts like placebo were known since XIX Century[5], the first clinical trial as we currently conceive it, was published in 1948[6]. The main feature of this modern approach is the respect to the experimentation subject, a need widely obviated by World War II[7].

Getting the most out of a medical scientific report is not a trivial task. A number of skills are required from the reader. Acquisition and development of those skills should start during undergraduate years and continue through one's professional life since methods have evolved closely as fast as Medicine itself.

Massive production of medical literature imposes a first challenge: selecting what is relevant or useless for a given reader has no rule-of-thumb. Also, quality can vary widely between reviews and within the same review. Furthermore, the mainstay of editors credibility, peer review, has been reported not to assure quality[8,9].

Quality and Bias

The very core of medical practice is decision making. What tests to order, building a list of differential diagnoses, figuring out the next question to ask, selecting the most appropriate therapy, which patient to care for with the highest priority are just some among the most remarkable. Real life situations situations are much different from the didactically organized theory of textbooks. Medical research articles are born in real daily practice. We read papers to make better decisions in the clinical setting, probably. And we should write papers in order to spread knowledge we produce.

The relevance of a research report is a function of how it influences clinical practice. Of course, this is a highly subjective definition in a general sense but certainly will make sense every time one uses or consider using a conclusion drawn from a research.

The randomized clinical trial (RCT), although seldom feasible in surgical investigation, is currently the gold standard in medical research when judging for quality. The reason is that RCTs are believed to provide means to control bias from several known sources, and perhaps some we are not aware of. Thus, the less undesired tendencies the more quality; the more of such tendencies the less quality.

Evidence-based Medicine

Although we can not assess quality or bias by the numbers, we can create a hierarchy based on the research methods that will most likely produce better or worse quality conclusions. This is the core of the Evidence-based Medicine (EBM).

There are many of such hierarchies. They vary among institutions[10] and, extremely important, if research regards treatment, prevention, diagnosis, prognosis or harm[11]. The choice of classification criteria is up to the reader and the institutions. Table 1 illustrates one possible hierarchy applicable to therapy research.

Table 1 – *Example of a hierarchy of quality of different study designs.*

Level	Type of study		
1	RCT; double-blind; placebo controlled	higher quality	less bias
2	other kinds or RCT		
3	non-randomized balanced controlled trials		
4	cohort or case-control analytic studies		
5	time series analysis		
6	descriptive studies (cases and series)		
7	reports of expert committees		
8	opinion of respected authority	lower quality	more bias

1. MD, SCT, Surgeon at Pirajussara General Hospital (Universidade Federal de São Paulo), consultant for Medical Research and Informatics.
2. Professor at Instituto de Matemática e Estatística da Universidade de São Paulo.

Recebido em 03/01/2008
Aceito para publicação em 15/02/2008

Conflito de interesses: nenhum
Fonte de financiamento: nenhuma

EBM emerged naturally from the availability of a reasonable amount of research, especially RCTs, the need for progressively better decisions and the need for standard institutional policies. The basic idea is to merge similar samples from several studies in the attempt to reach stronger conclusions by analyzing the pooled samples, the meta-analysis.

Although EBM has the reputation of highest reliability, it has some limitations[12]. Extrapolating to populations different from those from which samples were taken is usually considered a lower level of evidence[13]. Failure to publish negative trials is a major problem that has been addressed by registers of clinical trials[14,15]. Treatment effectiveness in the trial set tends to be higher due the closer follow up that enhances compliance rates[16-18].

At the individual decision level, EBM seems weaker[19] but there is growing evidence that it improves health care when applied at the institutional level as guidelines[20].

The reader must be aware that a lower quality report does not mean a bad article at all. It expresses only how directly one can apply its conclusions to daily practice. Actually, the trend of rejecting cases and series reports, both by editors and by readers, means that EBM's concept of quality has been widely misunderstood[21]. In EBM, only those studies ruled as well designed are taken into account. Thus, well designed and well analyzed studies of any kind will be considered of better quality than a poorly designed RCT. Cases and series reports are still good sources of information and discoveries that may guide future research and make the practitioner aware of previously unknown facts[21].

Evidence derived from meta-analysis must be stratified according to its strength represented by categories of recommendation. A sensible and widely used categorization is the one proposed by the U.S. Preventive Services Task Force[22].

Producing Evidence

The steps towards evidences are straightforward but arduous and tricky in the details. The work starts with a thorough review of the literature. All known sources of references must be included. Search criteria must be wide. This results in extensive lists but there is a smaller risk of missing some relevant piece of information. Lists are then scrutinized and possibly relevant articles are retrieved. All of them are read and the ones eligible to the study by previously defined inclusion criteria are selected for analysis[23]. This set of procedures is known as systematic review.

Initial lists may include thousands of references. Frequently more than a hundred papers need to be retrieved. A few of them will actually be relevant according to criteria like those in Table 1.

To avoid selection bias, the whole search and selection work is performed by a team of not less than three reviewers and usually all of them must agree that a given paper meets the inclusion criteria so it is actually included in the analysis[23].

Analysis

The samples extracted from each article are then evaluated for homogeneity. They must be similar so they can be compared or merged. Homogenous samples are then merged when possible and analyzed according to the questions the study intends to answer.[23] Some tools are typically used in EBM.

Likelihood Ratios

Derived from 2x2 contingency tables, the Likelihood Ratios (LRs) values may look strange to the reader since they do not lie in the unit interval like probabilities do. However, they are powerful tools for different sets. Using Bayesian statistical approaches, one can multiply the so-called "a priori" (or pre-test) probability by the likelihood ratio and derive directly the probability of disease (or non-disease),[24] the so-called "a posteriori" (or post-test) probability. Likewise, LRs can be multiplied by odds to produce the posterior odds. Odds and probabilities can easily be converted one to the other.

Suppose a surgeon works in a reference center for patients that are supposed to need emergency surgeries. This surgeon knows that about 60% of all abdominal pain cases will be operated on for appendicitis. He uses this number as his "a priori" probability (0.6) or odds (3/2) and multiplies by LR (disease/non-disease) of signs and symptoms previously determined relevant by EBM analysis and gets the "a posteriori" probabilities or posterior odds of disease/non-disease. Another surgeon in an emergency department is likely to have a smaller "a priori" probability, say 0.2 (20%) or odds 1/4. Despite the wide difference in the "a priori" probability in these examples it is so unlikely that it will impact the "a posteriori" probability or odds to the point of missing a significant proportion of diagnoses. Actually, chaining several LRs of several signs and symptoms as real situations demand has the reputation of overestimating the "a posteriori" probability.[24] In Surgery this feature may be highly desirable. This kind of decision rule has been efficient in the delivery of better health care even in critical situations[25-29].

There is a proposed fully Bayesian meta-analysis method that besides the probability of disease provides the strength of the evidence of each sign, symptom or test and the average strength of evidence, called diagnosability. This method has not been widely applied but has been shown to be a potentially powerful tool to produce desirable decision rules[30].

Area under the ROC Curve

A tool widely used in EBM is the receiver operator characteristic (ROC) curve. It represents the relationship between specificity and sensitivity. The area under the ROC curve expresses the power of a given test, sign or symptom to differentiate disease and non-disease states. The higher the values of sensitivity and specificity, the larger the area under the ROC curve, consequently, the better the test. The ideal test should produce an area equal to 1.

Another desirable product of the ROC curve is the cut-off value, the point of the curve that is closer to the upper left corner of the graph (Figure 1)[31] and expresses the best combination of sensitivity and specificity.

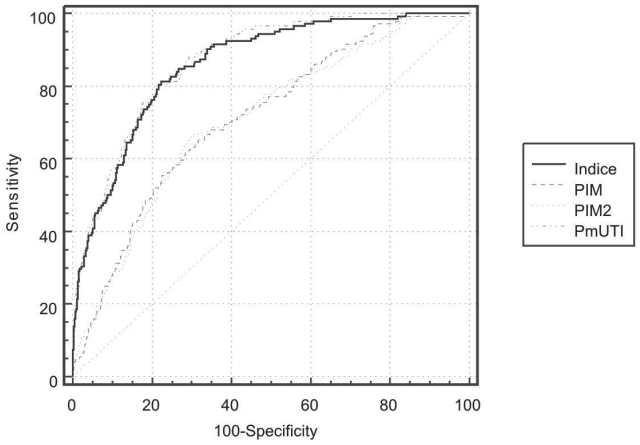

Figure 1 *– ROC curves of four methods designed to predict mortality in pediatric intensive care units.*[31]

High quality reports about tests or signs should provide information about the areas under the ROC curves as well as cut-off values for positive and negative tests or signs.

Number Needed to Treat

The number needed to treat (NNT) is expression of the intervention effectiveness. It can be applied to any treatment modality or diagnostic test. The proper way to reason with NNT is how many patients should be treated to have one positive response or, in the case of a test, how many tests should be performed to produce one diagnosis[32,33].

Many reports provide NNTs. From those that do not, the reader can easily calculate them from reported proportions[33].

The ideal NNT equals 1, i.e., every treatment will produce a positive response and every test will produce a diagnosis. NNTs of 2 or 3 represent a quite high effectiveness of an intervention or test. Most NNTs seen in clinical setting with effective interventions will range between 5 and 20[32,34-36]. The range from 20 to 40 may still represent clinical usefulness.[33] One must be aware that different NNTs, obtained in different clinical situations, should never be compared[36].

Number Needed to Harm

The number needed to harm (NNH) describes the number of patients we should treat to produce one more undesirable effect comparing to the alternative treatment,[37] ideally placebo.

It is intuitively apparent that the lesser the NNH, the worse the treatment. Besides drugs side effects, NNH can be used in cost analysis, for instance, but this kind of use deserve more sophisticated statistical methods[33].

There are no fixed limits to NNH. The risk of the clinical use of a treatment can only be properly assessed when we know the benefit this treatment produces in relation to the severity of illness.

Quality and Relevance

The primary purpose of medical practice is the delivery of good heath care. This is true for both the administrator and the physician. Although the latest method of research, powerful enough to produce wide and positive institutional influence and invaluable as a source of knowledge, EBM is not the definitive solution to questions the arise from clinical reasoning. For the daily work of the practitioner it is once more source of clues, as relevant as RCTs, and reports of lower strength from the EBM stand point. The concept of quality applies only for the EBM methods. For surgeons, physicians, practitioners, all sources of knowledge may have their relevance.

Final Remarks

After the first challenge, selecting good sources of knowledge, our second hard task regarding the vast literature is to define which acquired knowledge should change or influence our practice. This decision is private to each physician. It depends primarily of reader's background knowledge and experience. Sometimes this background is not enough for one to learn all a report can deliver.

This is the third hard task related to reading medical scientific reports: recognizing the need for and acquiring an adequate background.

Reading a paper has become a highly specialized skill. The ability to criticize and assess the relevance of any piece of medical information, always with a moderate dose of skepticism, is nowadays critical for any practitioner to be up to date.

Glossary

odds: probable number of times an event is likely to occur, expressed as the ratio of number of probable occurrences to the number of probable non-occurrences (source: Business Dictionary).

unit interval: the numeric interval between zero and one.

Bayesian: paradigm used to produce strictly probabilistic inferences, opposed to classical statistics that is based upon repetition of ideal experiments.

Conversion Rules and Calculations

P(d) – probability of disease

P(nd) – probability of non-disease

LR(d) – likelihood ratio of disease

LR(nd) – likelihood ratio of non-disease

O – odds

Op – posterior odds

P(d) = LR(d) * ["a priori" probability]

P(nd) = 1 – P(d)

Op = LR(d) * O

P = O / (1+O)

Examples

Take the odds 1/4 used as illustration above. The "a priori" probability of disease is given by

P(d) = O / (1+O) = 1 / (1+4) = 0.2 (20%)

and of non-disease is given by

P(nd) = 1 – P(d) = 0.8 (80%).

After processing a hypothetical decision rule for the diagnosis of appendicitis in patients with abdominal pain, the

surgeon obtained, for a given patient, LR(d) = 4.5 and LR(nd) = 0.125. Thus,

P(d) = 4.5 * 0.2 = 0.9 (90%) (this is the "a posteriori" probability of disease)

P(nd) = 0.125 * 0.8 = 1 – P(d) = 0.1 (10%) (this is the "a posteriori" probability of non-disease)

Besides,

Op_1 = 4.5 * 1 = 4.5 (posterior odds of disease)

Op_2 = 0.125 * 4 = 0.5 (posterior odds of non-disease)

Op = 4.5/0.5 = 9/1 (posterior odds)

Suggested Reading

Trisha Greenhalgh. How to Read a Paper: The Basics of Evidence-based Medicine. BMJ Books. London, UK, 2001. ISBN 0-7279-1578-9.

Acknowledgements

The authors are indebted to Dr. Lisieux Eyer de Jesus who provided insightful thoughts and invaluable suggestions to the text.

ABSTRACT

*Evidence-based Medicine (EBM) has become a major source of medical knowledge. It handles complexities of virtually every method or technique used in research. The knowledge on how the EBM researcher retrieves information, judges for relevance and analyzes derived data is invaluable for the skillful reader of medical scientific reports (*Rev. Col. Bras. Cir. 2008; 35(2): 141-145*).*

REFERENCES

1. Civil M. [Sumerian medical prescriptions]. Rev Assyriol Archeol Orient. 1960; 54:57-72.
2. Civil M. [A new Sumerian medical prescription]. Rev Assyriol Archeol Orient. 1961; 55:91-4.
3. Stol M. Blindness and night-blindness in Akkadian. J Near East Stud. 1986; 45(4):295-9.
4. Reynolds EH, Kinnier Wilson JV. Stroke in Babylonia. Arch Neurol. 2004; 61(4):597-601.
5. McQuay HJ, Moore RA. Placebo. Postgrad Med J. 2005; 81(953):155-60.
6. Medical Research Council Investigation. Streptomycin treatment of pulmonary tuberculosis. BMJ. 1998; 317:1248. Reprinted from: BMJ. 1948; 2:769-782.
7. Hedfors E. Medical ethics in the wake of the Holocaust: departing from a postwar paper by Ludwik Fleck. Stud Hist Philos Biol Biomed Sci. 2007; 38(3):642-55.
8. Richards D. Little evidence to support the use of editorial peer review to ensure quality of published research. Evid Based Dent. 2007; 8(3):88-9.
9. Jefferson T, Rudin M, Brodney Folse S, Davidoff F. Editorial peer review for improving the quality of reports of biomedical studies. Cochrane Database Syst Rev. 2007;(2):MR000016.
10. Elstein AS. On the origins and development of evidence-based medicine and medical decision making. Inflamm Res. 2004; 53 Suppl 2:S184-9.
11. Oxford Centre for Evidence-based Medicine. Levels of Evidence and Grades of Recommendation. http://www.cebm.net/index.aspx?o=1047. Retrieved on 09-03-2008.
12. Tonelli MR. The limits of evidence-based medicine. Respir Care. 2001; 46(12): 1435–40; discussion 1440-1.
13. Task Force Ratings. http://www.ahrq.gov/clinic/3rduspstf/ratings.htm. Retrieved on 09-03-2008.
14. Surgery Journal Editors Group. Consensus statement on mandatory registration of clinical trials [editorial]. J Ped Surg. 2007;42(4):601-2.
15. DeAngelis CD, Drazen JM, Frizelle FA, Haug C, Hoey J, Horton R, Kotzin S, et al. Clinical Trial Registration. A Statement from the International Committee of Medical Journal Editors [editorial]. JAMA. 2004; 292(11):1363-4.
16. Frolkis JP, Pearce GL, Nambi V, Minor S, Sprecher DL. Statins do not meet expectations for lowering low-density lipoprotein cholesterol levels when used in clinical practice. Am J Med. 2002;113(8):625-9.
17. Third Report on National Cholesterol Education Program (NCEP) expert panel on detection, evaluation and treatment of high blood cholesterol in adults (Adult Treatment Panel III): final report. IX. Adherence. Circulation. 2002; 106(25):3359-66.
18. Wei L, Wang J, Thompson P, Wong S, Struthers AD, MacDonald TM. Adherence to statin treatment and readmission of patients after myocardial infarction: a six year follow up study. Heart. 2002; 88(3):229-33.
19. Coomarasamy A, Khalid KS. What is the evidence that postgraduate teaching in evidence based medicine changes anything? A systematic review. BMJ. 2004; 329(7473):1017.
20. Yealy DM, Auble TE, Stone RA, Lave JR, Meehan TP, Graff LG, Fine JM, Obrosky DS, Mor MK, Whittle J, Fine MJ. Effect of increasing the intensity of implementing pneumonia guidelines: a randomized, controlled trial. Ann Intern Med. 2005; 143(12):881-94.
21. Neely JG, Karni RJ, Nussenbaum B, Paniello R, Fraley PL, Wang EW, Rich JT. Practical guide to understanding the value of case reports. Otolaryngol Head Neck Surg. 2008; 138(3):261-4.
22. United States Preventive Services Task Force. Task Force Ratings. http://www.ahrq.gov/clinic/3rduspstf/ratings.htm. Retrieved on 09-03-2008.
23. Sackett DL, Richardson WL, Rosenberg W. Evidence-base medicine: how to practice and teach EBM. New York: Churchill-Livingstone; 1997.
24. Hayden SR, Brown MD. Likelihood ratio: a powerful tool for incorporating the results of a diagnostic test into clinical decision making. Ann Emerg Med. 1999; 33(5):575-80.
25. Hess EP, Wells GA, Jaffe A, Stiell IG. A study to derive a decision rule for triage of emergency department patients with chest pain: design and methodology. BMC Emerg Med. 2008; 8:3.
26. Stiell IG, McKnight RD, Greenberg GH, McDowell I, Nair RC, Wells GA, Johns C, Worthington JR. Implementation of the Ottawa ankle rules. JAMA. 1994; 271(11):827-32.
27. Stiell IG, Wells GA, Hoag RH, Sivilotti ML, Cacciotti TF, Verbeek PR, Greenway KT, McDowell I, Cwinn AA, Greenberg GH. Implementation of the Ottawa knee rule for the use of radiography in acute knee injuries. JAMA. 1997; 278(23):2075-9.

28. Stiell IG, Wells GA, Vandemheen KL, Clement CM, Lesiuk H, De Maio VJ, Laupacis A, Schull M, McKnight RD, Verbeek PR, et al. The Canadian C-spine rule for radiography in alert and stable trauma patients. JAMA. 2001; 286(15):1841-8.
29. Wears RL, Li S, Hernandez JD, Luten RC, Vukich DJ. How many myocardial infarctions should we rule out? Ann Emerg Med. 1989; 18(9):953-63.
30. Pereira CAB, Perichi LR. Analysis of diagnosability. J R Stat Soc Ser C Appl Stat. 1990; 39(2):189-204.
31. Mangia CMF, Aguiar AS, Pereira CAB. Unpublished data.
32. Laupacis A, Sackett DL, Roberts RS. An assessmnet of clinically useful measures of the consequence of treatment. N Eng J Med. 1988; 318(26):1728-33.
33. Cordell WH. Number needed to treat (NNT). Ann Emerg Med. 1999; 33(4):433-6.
34. Rowe BH, Spooner CH, Ducharme FM, Bretzlaff JA, Bota GW. The effectiveness of corticosteroids in the treatment of acute exacerbations of asthma: a meta-analysis of their effect on relapse following acute assessment. Cochrane Database of Systematic Reviews. 1998; 3:3.
35. Tissue plasminogen activator for acute ischemic stroke. The National Institute of Neurological Disorders and Stroke rt-PA Study Group. N Eng J Med. 1995; 333(24):1581-7.
36. McQuay HJ, Moore RA. Using numerical results from systematic reviews in clinical practice. Ann Intern Med. 1997; 126(9):712-20.
37. Sackett DL, Haynes RB. Summarising the effects of therapy: a new table and some more terms. [EBM Note] Evidence-Based Medicine. 1997; 2:103-4.

Como citar este artigo:
Aguiar AS, Pereira CA. Weight of evidence. Rev Col Bras Cir. [periódico na Internet] 2008; 35(2). Disponível em URL: http://www.scielo.br/rcbc

Correspondence Address
Alexandre Santos Aguiar
email: asaguiar@spsconsultoria.com
R Botucatú, 591 cj 81
04023-062 - São Paulo – SP